Transactions of the International Society for
Music Information Retrieval

Ycart, A., Liu, L., Benetos, E., and Pearce, M.T. (2020). Investigating the Perceptual Validity of Evaluation Metrics for Automatic Piano Music Transcription, *Transactions of the International Society for Music Information Retrieval*, V(N), pp. xx–xx, DOI: https://doi.org/xx.xxxx/xxxx.xx

RESEARCH ARTICLE

Investigating the Perceptual Validity of Evaluation Metrics for Automatic Piano Music Transcription

Adrien Ycart*, Lele Liu*, Emmanouil Benetos*, and Marcus T. Pearce*†

Abstract

Automatic Music Transcription (AMT) is usually evaluated using low-level criteria, typically by counting the numbers of errors, with equal weighting. Yet, some errors (e.g. out-of-key notes) are more salient than others. In this study, we design an online listening test to gather judgements about AMT quality. These judgements take the form of pairwise comparisons of transcriptions of the same music by pairs of different AMT systems. We investigate how these judgements correlate with benchmark metrics, and find that although they match in many cases, agreement drops when comparing pairs with similar scores, or pairs of poor transcriptions. We show that onset-only notewise F-measure is the benchmark metric that correlates best with human judgement, all the more so with higher onset tolerance thresholds. We define a set of features related to various musical attributes, and use them to design a new metric that correlates significantly better with listeners' quality judgements. We examine which musical aspects were important to raters by conducting an ablation study on the defined metric, highlighting the importance of the rhythmic dimension (tempo, meter). We make the collected data entirely available for further study, in particular to evaluate the perceptual relevance of new AMT metrics.

Keywords: Automatic music transcription; polyphonic music similarity; evaluation metrics.

1. Introduction

Automatic Music Transcription (AMT) is a widely discussed problem in Music Information Retrieval (MIR) (Benetos et al., 2019). Its ultimate goal is to convert an audio signal into some form of music notation, such as sheet music, which we refer to as Complete Music Transcription (CMT). A common intermediate step is to obtain a MIDI-like representation, describing notes by their pitch, onset and offset times in seconds, leaving aside problems such as stream separation, rhythm transcription, or pitch spelling. We refer to this as AMT. It has applications in various fields, in particular in music education, music production and creation, musicology, and as pre-processing for other MIR tasks, such as cover detection or structural segmentation.

The performance of AMT systems is commonly assessed using simple, low-level criteria, such as by counting the number of mistakes in a transcription (Bay et al., 2009). In particular, deciding whether a note is a mistake is typically a binary decision, and all errors have the same weight in the final metric. Yet, not all mistakes are equally salient to human listeners: for instance, an out-of-key false positive will be much more noticeable than an extra note in a big chord, all the more so if it fits with the harmony.

In this study, we aim to investigate to what extent the current evaluation metrics correlate to human perception of the quality of an automatic transcription. We reframe the problem of AMT evaluation as a symbolic music similarity problem: we try to assess how similar to the target the output transcription sounds, rather than simply counting the number of incorrectly detected notes. We gather judgements of similarity by conducting a listening test, and use these answers to examine how human perception of AMT quality correlates with the evaluation metrics commonly used. We investigate what musical features are most important to raters, and use them to define a new metric, that correlates significantly better with human ratings than benchmark metrics.

Gathering similarity ratings in a meaningful way is not straightforward. In particular, inter-rater agreement is infamously low for music similarity tasks (Flexer and Grill, 2016). One of the reasons, besides intrinsic disagreement between raters, is that it is a difficult and ill-defined task. Our main concern is thus to make the test as easy as possible. As argued by Al-

*Centre for Digital Music, School of Electronic Engineering and Computer Science, Queen Mary University of London, UK

†Center for Music in the Brain, Aarhus University, Denmark

lan et al. (2007), the difficulty of rating the absolute similarity between two excerpts, be it on a continuous or Likert scale (Likert, 1932), leads to low inter-rater agreement, as different raters might use different scales, and these scales might evolve throughout the experiment. To avoid that problem, we choose to give raters a binary choice: given one reference excerpt, and two possible transcriptions of that excerpt, participants have to answer the question "Which transcription sounds most similar to the reference?". Another reason that makes rating difficult is having to remember long excerpts for subsequent comparison. In order to make the task easier, such that participants can rely mostly on their working memory, we use short audio excerpts, which prevents us from drawing any conclusions on the similarity of longer excerpts. Since we are mostly interested in notes rather than timbre or sound quality, we can afford to run this study in more loosely controlled acoustic conditions. We thus run this study online, in order to gather as much data as possible. A major concern is to make the test easily accessible; in particular, it is designed so participants can answer as many or as few questions as they want.

We choose to focus our study on Western classical piano music, as it is by far the most discussed subdomain of AMT, mostly due to the availability of big datasets for that instrument and style (Emiya et al., 2010; Hawthorne et al., 2019). The validity of the present study is thus limited to this instrument and style, and should not be generalised e.g. to singing voice, or jazz music.

Our main contributions include:

- Gathering a dataset of more than four thousand individual perceptual ratings of transcription quality;
- Investigating the correlation between these ratings and traditional AMT metrics, depending on various factors;
- Proposing a set of musically-relevant features that can be computed on pairs of target and AMT output;
- Proposing a new evaluation metric in the form of a simple logistic regression model trained to approximate listener ratings;
- Investigating which musical parameters are most important to raters through an ablation study of the classifier.

In particular, we make the stimuli, gathered data, website code, pre-trained metric and feature implementation all available for further study (See Section 7).

In what follows, we present the benchmark evaluation metrics used for AMT and other works on transcription system evaluation in Section 2, and describe the design of the listening tests in Section 3. In Section 4, we analyse the results of the listening tests, and in particular the agreement between ratings and benchmark evaluation metrics. We then define a new metric based on musical features and analyse which features were most important to users in Section 5. Finally, we discuss our results in Section 6.

2. Related works

2.1 Benchmark evaluation metrics

In this section we describe the most commonly-used evaluation metrics for AMT. Some other metrics exist (see Bay et al. (2009) for a complete description), we only briefly describe here those that are most often used to compare systems.

2.1.1 Framewise metrics

These metrics are computed on pairs of piano rolls. A piano roll is a binary matrix M, such that $M[p,t] = 1$ if and only if pitch p is active at frame t, where a frame is a temporal segment of constant duration. We use a timestep of 10ms, as in the MIREX multiple-F0 estimation task (Bay et al., 2009). When comparing an estimated piano roll $\hat{M}$ to a target piano roll M, a true positive is counted whenever $\hat{M}[p,t] = 1$ and $M[p,t] = 1$. False positives and false negatives are counted analogously. We use TP, FP and FN to refer to the total number of true positives, false positives and false negatives, respectively, summed across frames.

The framewise Precision (P_f), Recall (R_f) and F-Measure (F_f) are then computed as follows (the subscript f represents the fact that metrics are computed framewise):

$$P_\mathrm{f} = \frac{TP}{TP+FP} \quad R_\mathrm{f} = \frac{TP}{TP+FN} \quad F_\mathrm{f} = \frac{2 \cdot P_\mathrm{f} \cdot R_\mathrm{f}}{P_\mathrm{f} + R_\mathrm{f}} \qquad (1)$$

2.1.2 Notewise metrics

Notewise metrics are computed on lists of notes, where each note is a tuple (s,e,p) where s and e are the start and end times, and p is the MIDI pitch of the note. For onset-only notewise metrics, an estimated note $(\hat{s},\hat{e},\hat{p})$ is considered as a true positive if and only if there is a ground-truth note (s,e,p) such as $p = \hat{p}$ and $|s - \hat{s}| <$ 50ms. Besides, ground-truth notes can be matched to at most one estimated note. Precision, Recall and F-Measure (respectively $P_\mathrm{n,On}$, $R_\mathrm{n,On}$ and $F_\mathrm{n,On}$) are then computed as in Section 2.1.1, with the difference that TP, FP and FN are counted in number of notes, instead of time-pitch bins. The subscript n represents the fact that metrics are computed notewise.

Recently, as $F_\mathrm{n,On}$ performance for AMT systems has improved, onset-offset notewise metrics have been increasingly used. Onset-offset metrics add the extra constraint that, for an estimated note to be considered a true positive, $\hat{e}$ must be within 20% of the duration of the ground-truth note or within ±50ms of the ground truth offset, whichever is greatest. Again, Precision, Recall and F-Measure (respectively $P_\mathrm{n,OnOff}$, $R_\mathrm{n,OnOff}$ and $F_\mathrm{n,OnOff}$) are computed as in Section 2.1.1.

In all cases, metrics are computed for each test piece, and then averaged over the whole dataset. In

particular, we do not weigh each piece according to its duration.

2.2 Efforts for better evaluation metrics

Recently, various evaluation methods were proposed for CMT (Cogliati and Duan, 2017; McLeod and Steedman, 2018), but they focus mostly on typesetting problems, and do not address the problem of perceptually-relevant pitch assessment. Some efforts were also made for singing voice transcription and melody estimation (Molina et al., 2014; Bittner and Bosch, 2019), but still consider pitches as being either correct or incorrect. Another method was proposed for automatic solfège assessment in (Schramm et al., 2016), using a classifier trained on experts ratings to classify each note as correct or incorrect, but again, this decision is mostly binary, and focuses on small deviations in pitch (less than a semitone) rather than the correctness of a pitch in a tonal context.

An older study was conducted on AMT by Daniel et al. (2008). The study assessed the perceptual discomfort created by some specific types of mistakes (e.g. note insertions, deletions, replacement, onset displacement) by comparing pairs of artificially-modified music excerpts. This data was then used to define new evaluation metrics. However, the types of mistakes considered were relatively limited (for instance, for note insertions, the study only compared octave insertions, fifth insertions and random insertions), and did not take into account musical concepts such as tonality, melody, harmony, or meter. Moreover, the modified MIDI files only contained one type of mistake, and did not consider the potential interactions between several kinds of mistakes. By contrast, we choose to use real AMT system outputs, in order to maintain ecological validity, and study a wider range of features.

The evaluation of AMT systems is related to symbolic music similarity, as the end goal is to assess how similar the output and the target sound. Symbolic melodic similarity is a widely-discussed problem (see Velardo et al. (2016) for a survey). Here, we are focusing on polyphonic music similarity, which is much less common. A method is described by Allali et al. (2009), relying on sequence-to-sequence alignment, and an edit distance adapted from Mongeau and Sankoff (1990). However, this method was designed for quantised note durations only, which makes it potentially suitable for CMT, but not for AMT. Moreover, we aim here to use a bottom-up approach, to investigate what factors are important to listeners and using them to define a new metric.

3. Study design

3.1 Stimulus design

We obtain automatic transcriptions using several benchmark AMT systems. Using the best systems available currently would have led to very similar transcription mistakes, as they are all based on the same underlying methods. Instead, we aim to use a diverse sample of commonly used AMT methodologies. We thus use:

OAF: The current state of the art based on neural networks (Hawthorne et al., 2019), trained to jointly detect note onsets and pitches.

CNN: A simple framewise convolutional neural network (Kelz et al., 2016).

NMF: A piano-specific system, based on non-negative matrix factorisation (Cheng et al., 2016).

STF: A system based on handcrafted spectral and temporal features (Su and Yang, 2015).

CNN is a framewise system: at each timestep, it outputs a list of active pitches. This is equivalent to a piano roll, but requires post-processing to obtain a list of note events. To get note events, we consider any silence followed by a note as an onset (and vice versa for offsets), and apply gap-filling and short-note-pruning, both with a threshold of 80ms, corresponding to two processing frames in this system.

We use the pieces present in the MAPS dataset (Emiya et al., 2010) of MIDI-aligned piano recordings, as it remains the most common benchmark dataset for AMT. We use only the full music pieces in MAPS, with the two recording conditions that correspond to real piano recordings, namely ENSTDkCl (close-field recordings) and ENSTDkAm (ambient recordings), the two most commonly-used evaluation subsets. To preserve musical validity, we manually segment the pieces into musical phrases, so that each excerpt lasts between five and ten seconds and roughly corresponds to a coherent, self-contained musical unit. We try as much as possible to keep an integer number of bars, using the A-MAPS (Ycart and Benetos, 2018) bar and beat annotations. When material within a piece is repeated without transposition, we only keep the first repetition. The start and end times of each segment are made available for future study (see Section 7). We keep duplicate pieces, recorded with two different recording conditions. Eventually, we obtain 1552 reference examples.

To be as consistent as possible in terms of timbre between the reference and the transcriptions, all example MIDI files were rendered using the Yamaha Disklavier Pro Grand Piano soundfont.[1] Some systems could not transcribe note velocities, so for uniformity, we used a default MIDI velocity of 100 for every note of the output transcriptions. We kept the original velocities when rendering references to be able to use them later on in the analysis, as most of the time they are available in the ground-truth files.

3.2 User data

Before answering questions, users read an information sheet and gave their consent for participating. We collected their age, gender, and whether they had a hearing disability. They then had to answer questions

from the Gold-MSI test (Müllensiefen et al., 2014) corresponding to the Perceptual Abilities and Musical Training subscales. Each user also had the option to give comments on the strategies they used and the aspects that were most important to them when choosing between transcriptions. All data was anonymised, and the procedure was approved by Queen Mary University of London's ethics committee (reference QMREC2066).

3.3 Setup

The test was conducted online, as the main focus of this study was not sound quality, but rather the note content of the transcriptions. Participants were advised to do the test using good headphones, in a quiet environment. In what follows, we call a question a set {reference,transcription1,transcription2}, where transcription1 and transcription2 are two transcriptions of the reference, made by two different systems. There are six questions per reference, one for each unordered pair of AMT systems. For each question, participants were presented with one "reference" audio player, two "transcription" audio players, and were asked to answer the question "Which transcription sounds most similar to the reference?", as a two-alternative forced choice (see Figure 1 for a screenshot of the interface). To strike a balance between comparison robustness and number of answered questions, each question was rated by four participants, taking care to balance the order (transcription1, transcription2) and (transcription2, transcription1) in which the two transcription players are presented in the interface. Participants were allowed to listen to each example as many times as they wanted; however, to encourage them to rely on perception rather than analytical thinking, we advised participants to listen to each example as few times as possible. A five-minute time limit was also included. For each question, participants could report if they knew the reference by ticking an additional "I know this piece" box.

While designing the test, it became apparent that in some instances, making a choice was very difficult, for instance when the two transcriptions were nearly identical, or different but equally poor. We did not want to include a third alternative (such as "I don't know", or "both transcriptions are equally similar to the target"), as this would have made it much more difficult to produce a meaningful analysis of the difficult cases. Instead, we added an extra question: "How difficult was it to answer the question?", on a five-point Likert scale (Likert, 1932) from "Very easy" to "Impossible". Guidelines were given to answer this question in terms of number of listenings required for each file, difficulty of making a choice, and confidence in that choice.

Getting participants to spend 30 minutes or more on a listening test without compensation can be difficult. To allow more flexibility, we designed the test so that each participant could rate as many examples as they wanted. If we had randomly picked questions, given the large number of examples, it would have been very difficult to ensure that several people answered each question. Instead, questions were presented to participants using the following rules:

1. Each participant cannot hear a reference more than once.
2. Each question cannot be rated more than four times.
3. Each new question is chosen among remaining candidates using the following steps:
   (a) Choose a reference among those that have already been seen by other participants, and have not been fully rated (*i.e.* at least one of the six questions using that reference has less than four answers).
   (b) If no such reference is available, choose a random new question.
   (c) Otherwise, choose a question using that reference that has already been answered by other participants.
   (d) If no such question is available, choose a new question using the same reference.

When choosing a reference among those that have been seen by other participants (step 3.(a)), we skewed the random choice towards references that had more answers, in order to maximise the number of fully-rated references (i.e. references for which all system pairs were rated by four participants). Thanks to this procedure, the size of the pool of examples adapted dynamically to the number of gathered answers.

3.4 Participants

In total, 186 people participated in our study (excluding the 40 people that registered but did not answer any questions): 126 males, 58 females and 2 non-binary, with a median age of 28. We did not make any selection on participants. Many of them were trained musicians, as the median Gold-MSI score is 5.06 on a scale from 1 to 7 (compared to 4.81 in the general population for the subscales considered (Müllensiefen et al., 2011)). The median number of answered questions was 20, with 22 participants answering 50 questions or more (up to several hundreds). Overall, we gathered 4501 answers, 1080 questions with four ratings, and 153 examples for which all pairs of systems have four ratings. Four participants reported a hearing disability, for a total of 53 answers. We decided to keep them anyway, as they amount for a small proportion of answers, and we are not interested in fine judgement about sound quality.

4. Results

In what follows, we analyse the results of the participants' ratings. We only keep questions for which four answers have been gathered. We keep all such ques-

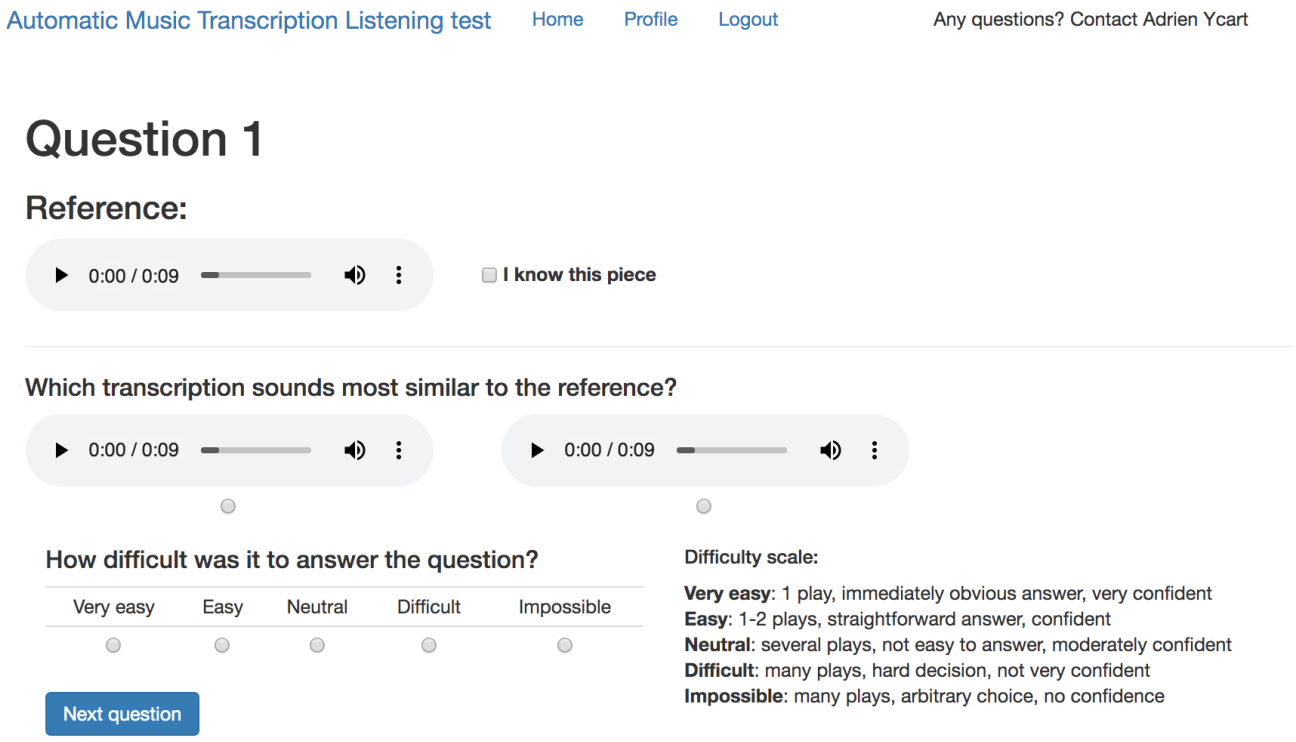

Figure 1: Screenshot of the listening test website.

tions, even when the corresponding example has not been rated for all pairs of systems. When comparing proportions (e.g. user preference, or agreement between raters and benchmark metrics), error bars are obtained by bootstrap analysis (Efron, 1992), resampling with same dataset size 100 times. The standard deviation of bootstrapped results is displayed.

4.1 Benchmark system performance

First, we run the chosen systems on all the test files. We evaluate them using the benchmark metrics described in Section 2.1. Results are presented in Table 1. Notewise metrics were computed using the `mir_eval` Python library (Raffel et al., 2014).

As expected, OAF is by far the best of all, for all metrics. The second-best is NMF, which can also be explained by the fact that is was trained on that specific instrument model, while this piano model is new to the other systems. The CNN comes in third position, and STF comes last.

It has to be noted that these results vary quite a lot between the two subsets ENSTDkCl and ENSTDkAm: results are usually worse on ENSTDkAm, since it corresponds to ambient piano recordings, which are usually noisier. In particular, for NMF, which was trained on isolated notes played on ENSTDkCl, $F_{\mathrm{n,On}}$ drops from 76.1 to 55.6 on ENSTDkAm. For CNN and STF, $F_{\mathrm{n,On}}$ drops of around 5%. Interestingly, OAF works similarly on both subsets. This can be explained by the fact that it was trained on the MAESTRO dataset (Hawthorne et al., 2019), a dataset containing mostly concert piano recordings, in conditions arguably closer to ENSTDkAm.

It also appears that although the performance in F_{f} is within a relatively small range of values, there are much bigger differences in performance in terms of $F_{\mathrm{n,On}}$ and $F_{\mathrm{n,OnOff}}$.

4.2 Perceptual ranking of systems

Using the ratings, we evaluate the systems from a perceptual point of view (pairwise results shown in Figure 2). The ratings are generally in accordance with the benchmark metrics: a system is preferred when its $F_{\mathrm{n,On}}$ is better (we focus on $F_{\mathrm{n,On}}$ as this metric correlates best with ratings, as discussed in Section 4.3). The relative ranking of the systems is also the same: OAF beats all other systems, NMF beats CNN and STF, and CNN beats STF. There seems to be a relation between the difference in benchmark metrics and the magnitude of the majority: for instance, OAF has a bigger majority when compared to STF than to NMF. But that is not strictly the case: although CNN is much better than STF in terms of $F_{\mathrm{n,On}}$ and $F_{\mathrm{n,OnOff}}$, it is only preferred about 65% of the time.

4.3 Agreement between ratings and benchmark metrics

In this section, we assess the extent to which ratings agree with F_{f}, $F_{\mathrm{n,On}}$ and $F_{\mathrm{n,OnOff}}$. We also investigate what factors influence the agreement between raters

System	P_{f}	R_{f}	F_{f}	$P_{\mathrm{n,On}}$	$R_{\mathrm{n,On}}$	$F_{\mathrm{n,On}}$	$P_{\mathrm{n,OnOff}}$	$R_{\mathrm{n,OnOff}}$	$F_{\mathrm{n,OnOff}}$
STF	67.2	60.0	62.7	49.8	32.0	38.3	16.5	11.3	13.2
CNN	80.2	58.2	66.1	77.0	54.9	63.2	33.5	24.6	28.0
NMF	71.3	63.3	66.4	79.6	57.0	65.7	35.7	26.4	30.0
OAF	**89.0**	**79.5**	**83.8**	**85.9**	**84.1**	**84.9**	**66.9**	**65.5**	**66.2**

Table 1: Benchmark evaluation metrics for all systems, evaluated on the MAPS subsets ENSTDkCl and ENSTDkAm, with best values in bold.

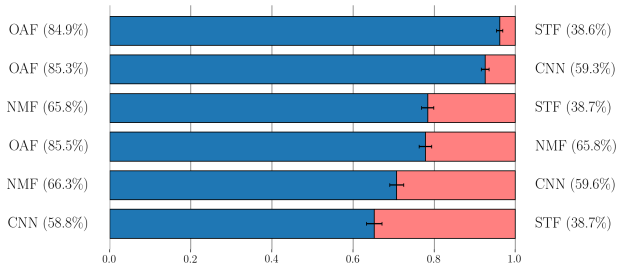

Figure 2: Vote proportion in pairwise comparisons of the systems. Blue bars represent the proportion of times the system on the left was chosen over the one on the right. For each pair, the percentage in parenthesis is the average $F_{\mathrm{n,On}}$ computed on the specific examples included in the comparison.

and benchmark metrics.

We define the agreement with a given metric as follows. For each given answer, we check whether the choice made by the participant corresponds to the ordering of the two transcriptions according to this metric. If the participant chose the transcription for which the metric is highest, we consider that the participant and the metric agree. We then compute the proportion of ratings that agree with this metric. We do as such for F_{f}, $F_{\mathrm{n,On}}$ and $F_{\mathrm{n,OnOff}}$. For F_{f}, we investigate various frame sizes: 10, 50, 75, 100, and 150ms. For notewise metrics, we investigate how this agreement varies depending on the onset and offset tolerance thresholds: for onsets, we use 25, 50, 75, 100, 125, and 150ms, and for offsets, we use 10, 20, 30, 40, and 50% of the note duration.

Results on the agreement between ratings and benchmark metrics are shown in Figures 3 and 4. In terms of frame size for F_{f}, there is no clear tendency. It does appear nonetheless that using a 100ms frame size improves slightly but significantly the agreement with ratings compared to a 10ms frame size ($p < 10^{-3}$ with a Welch t-test). When examining the influence of the onset for $F_{\mathrm{n,On}}$, we can see in Figure 3 that the agreement with ratings is highest for $F_{\mathrm{n,On}}$, for onset thresholds between 75 and 150ms. For $F_{\mathrm{n,OnOff}}$, we can see in Figure 4 that the agreement is highest for an onset threshold of 100ms and an offset tolerance of 50%, although it is still lower than $F_{\mathrm{n,On}}$ with onset threshold above 50ms. Agreement might be even higher for higher offset tolerance thresholds, as $F_{\mathrm{n,OnOff}}$ becomes more and more similar to $F_{\mathrm{n,On}}$ ($F_{\mathrm{n,On}}$ can be seen as $F_{\mathrm{n,OnOff}}$ with an infinite offset tolerance).

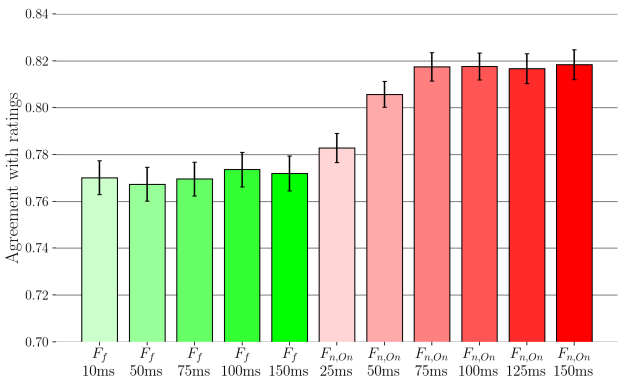

Figure 3: Percentage of agreement, across all examples, between raters and various evaluation metrics (F_{f} with various frame sizes, and $F_{\mathrm{n,On}}$ with various tolerance thresholds).

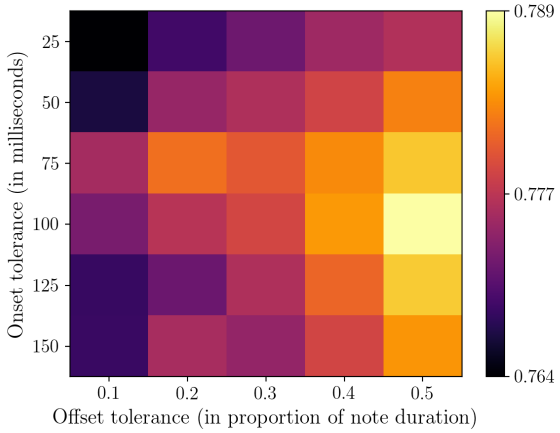

Figure 4: Percentage of agreement, across all examples, between raters and $F_{\mathrm{n,OnOff}}$, with various onset and offset tolerance thresholds.

To investigate further what factors might influence agreement, we perform a linear mixed effects analysis (Baayen et al., 2008), using as the dependent variable for each question whether the rater agreed with $F_{\mathrm{n,On}}$ (1 if they do, 0 otherwise). We use as fixed effects the best $F_{\mathrm{n,On}}$ of the pair (F_{best}), the difference in $F_{\mathrm{n,On}}$ between the two transcriptions (ΔF), the Gold-MSI score

of the rater (Gold-MSI), whether the piece was recognised (Known), and the reported difficulty (Difficulty). We use no random effects. The resulting coefficients and associated p-values are given in Table 2.

Feature	Coefficient	P-value
ΔF	0.539	<0.001
F_{best}	0.330	<0.001
Gold-MSI	-0.007	0.232
Known	0.014	0.391
Difficulty	-0.044	<0.001

Table 2: Coefficients and p-values for the linear mixed effects model using agreement with $F_{\text{n,On}}$ as dependent variable and features as fixed effects.

It appears that ΔF and F_{best} have a strong and significant effect on agreement. When the difference in performance between the two systems is high, people tend to agree more with the F-measure, as the choice is clearer. However, for a given ΔF, when both systems produce outputs of poor quality, the agreement is lower.

When looking at other features, Difficulty is negatively correlated with agreement: when people report the choice as being more difficult, they tend to disagree more with F-measure. To investigate this further, we compute the percentage of agreement between ratings and F-measure for each reported difficulty level (Figure 5). For high levels of difficulty, agreement is very poor, close to chance (50% for a two-alternatives forced choice question), which is consistent with the guidelines given to raters for reporting difficulty. Still, even for low levels of reported difficulty, there is a fair amount of disagreement between ratings and $F_{\text{n,On}}$ (10 to 20%), which shows that disagreement with $F_{\text{n,On}}$ does not exclusively result from random choices in the difficult cases. Musical training (Gold-MSI) and familiarity (Known) have no significant effect on agreement with $F_{\text{n,On}}$.

4.4 Reported difficulty

In this section, we examine the reported level of difficulty for each answer, and investigate the factors that influenced it.

In Figure 6, we display the proportion of ratings for each difficulty level. When comparing this figure to the results in Table 1, it appears that, as a general trend, the higher the difference in $F_{\text{n,On}}$, the more confident raters are. Moreover, difficulty is highest when comparing the two worst performing systems according to benchmark metrics, which suggests that difficulty is higher when both transcriptions are poor.

To get a better understanding of how the difficulty varies depending on various parameters, we perform another linear mixed effects analysis, using this time difficulty as dependent variable. We use as fixed effects the best $F_{\text{n,On}}$ of the pair (F_{best}), the difference in

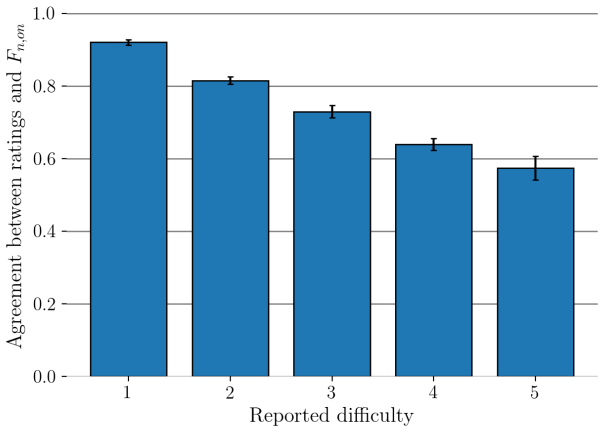

Figure 5: Agreement between ratings and $F_{\text{n,On}}$ for each reported difficulty level

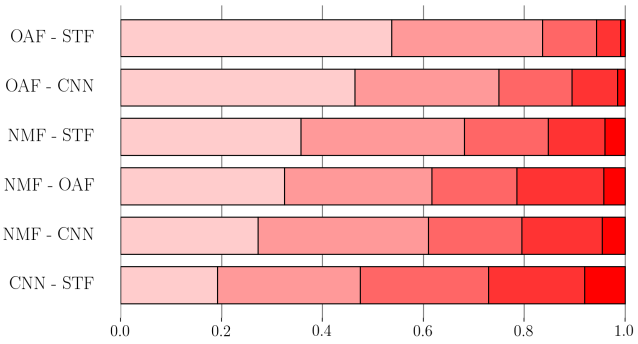

Figure 6: Proportion of difficulty ratings (lightest=1, darkest=5) for each pair of systems.

$F_{\text{n,On}}$ between the two transcriptions (ΔF), the Gold-MSI score of the rater (Gold-MSI), whether the piece was recognised (Known), and whether the rater agreed with $F_{\text{n,On}}$ (Agree). Again, we use no random effects. The resulting coefficients and associated p-values are given in Table 3.

All of the these factors are significant predictors of reported difficulty. From this, we can draw the following conclusions. First, musicians found the task easier than non-musicians. This could be explained either in terms of better auditory skills, or because musicians tend to be more confident in their judgements. People also find it easier to make a choice when they know the reference. One user commented: "Songs that I knew already felt easier to judge as I could remember the original much better", in other words they only had to listen to and remember two excerpts instead of three. This highlights a difficulty of investigating musical similarity perception due to effects of memory, as we mentioned in Section 1. It also appears that the more confident people are in their choices, the more they agree with F-measure, which is coherent with the results presented in Section 4.3. Finally, when investigating the effect of ΔF and F_{best}, we can see that the bigger the

Feature	Coefficient	P-value
ΔF	-1.564	<0.001
F_{best}	-0.608	<0.001
Gold-MSI	-0.227	<0.001
Known	-0.153	0.002
Agree	-0.423	<0.001

Table 3: Coefficients and p-values for the linear mixed effects model using difficulty as dependent variable and features as fixed effects.

Feature	Coefficient	P-value
ΔF	0.584	<0.001
F_{best}	349	<0.001
Gold-MSI	-0.014	0.011
Known	0.002	0.912
Difficulty	-0.036	<0.001

Table 4: Coefficients and p-values for the linear mixed effects model using agreement with $F_{\text{n,On}}$ as dependent variable and features as fixed effects, on confident answers only.

difference between the two systems, the easier the decision, and all the more so when both systems perform well.

4.5 Analysis of confident answers

When discussing the agreement between ratings and $F_{\text{n,On}}$, it is not straightforward to distinguish cases when participants chose randomly from cases where they actually disagreed with $F_{\text{n,On}}$, in particular where the two options have similar $F_{\text{n,On}}$, or when both options are poor. To avoid cases of random choice, we analyse the subset of answers that are confident (Difficulty=1 or 2, which represents 2856 answers), and investigate whether different factors influence the agreement between ratings and $F_{\text{n,On}}$ in this case.

We perform the same linear mixed effect analysis as in Section 4.3, on that subset. The results are shown in Table 4 and are quite similar to the full analysis, except that now there is a significant negative correlation between Gold-MSI and agreement. For confident answers, it appears that musicians tend to disagree more with $F_{\text{n,On}}$ than non-musicians. This could indicate that musicians focus more on certain high-level aspects of the music (e.g. melody, harmony, meter) that are not taken into account by $F_{\text{n,On}}$: even if it contains more mistakes, a transcription might be preferred by a musician as long as it gets these aspects right.

When investigating the effect of the difference in $F_{\text{n,On}}$ on agreement, we see once again the same trend: the smaller the difference between the two transcriptions, the greater the disagreement, as shown in Figure 7. When the difference in $F_{\text{n,On}}$ is above 50%, people always agree with $F_{\text{n,On}}$. However, below this threshold, agreement declines, especially when the difference is below 20%.

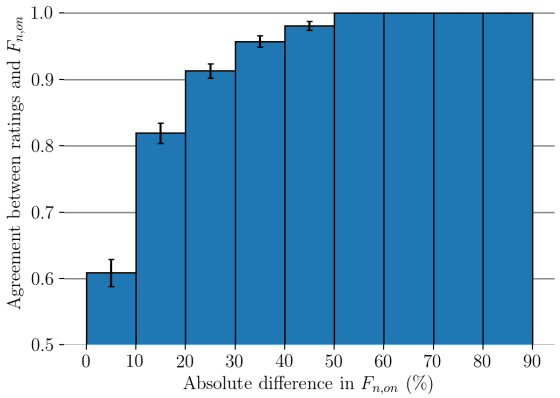

Figure 7: Percentage of agreement depending on the difference in $F_{\text{n,On}}$ between the two options, computed on confident answers only.

4.6 Inter-rater agreement

We have seen that there is a fair amount of disagreement between the F-measure and ratings. To get an idea of how consistent the ratings are, we investigate the level of inter-rater agreement, and the factors that influence it.

We begin by computing Fleiss' Kappa coefficient (Fleiss, 1971), that represents inter-rater agreement for an arbitrary number of raters. When computed over the whole dataset, we obtain a Kappa coefficient of 0.59, which can be interpreted as borderline between moderate and substantial agreement. When computing the same coefficient on the confident answers only (keeping only questions for which four confident answers were given, 315 questions in total), we obtain a Kappa coefficient of 0.90, which can be interpreted as near-perfect agreement. This is a very conservative estimate, as we keep only the questions that were unanimously considered as easy to answer. Moreover, inter-rater agreement is high because most of the time, raters tend to agree with F-measure.

We run a linear mixed effect analysis using the amount of agreement between raters as dependent variable (2 if all four raters agree, 1 if one rater disagrees with the other three, and 0 in the case of a draw), only on the subset of confident answers, and keeping only the questions with four confident answers. We use as dependent variables the difference of $F_{\text{n,On}}$ between the two systems (ΔF), the best $F_{\text{n,On}}$ of the pair (F_{best}), the average and standard deviation of the Gold-MSI scores of the four raters for each question (Gold-MSI$_{\text{avg}}$ and Gold-MSI$_{\text{std}}$ respectively), and the average reported difficulty (Difficulty$_{\text{avg}}$). The resulting coefficients and associated p-values are given in Table 5.

Once again, we observe that the bigger the differ-

Feature	Coefficient	P-value
ΔF	0.496	<0.001
F_{best}	-0.092	0.423
Gold-MSI$_{\text{avg}}$	-0.071	0.004
Gold-MSI$_{\text{std}}$	-0.016	0.778
Difficulty$_{\text{avg}}$	-0.176	0.003

Table 5: Coefficients and p-values for the linear mixed effects model using agreement among raters as dependent variable and features as fixed effects.

ence in $F_{\text{n,On}}$, the higher the agreement among raters. However, this time, the $F_{\text{n,On}}$ of the best solution does not seem to have a significant effect (noting that we also have many fewer data points). Raters also tend to disagree more with each other when the reported difficulty is higher on average. It also appears that when raters have a high average Gold-MSI, they tend to disagree more with each other. This could be due to the fact that trained musicians might favour different aspects of music (rhythm rather than melody for instance) when making a choice. Disparity in Gold-MSI among raters has no significant effect on whether they agree.

4.7 Discussion

It appears that the best correlation with ratings is achieved for much higher tolerance thresholds than what is usually used for transcription system evaluation, both for $F_{\text{n,On}}$ and $F_{\text{n,OnOff}}$. This suggests people are generally relatively forgiving with respect to onset precision, and probably focus on other aspects of music than just onset and offset precision to make their choices. Moreover, the OnOff-Note metric, presented as the most perceptually-relevant evaluation metric by Hawthorne et al. (2018), is actually not the best metric in terms of agreement with human ratings, at least in the case of piano music. On-Note metrics should be favoured, though this may relate to the focus on piano which generally has very salient onsets, but less clear offsets, especially for long notes. OnOff-Note metrics are still useful from an engineering perspective, as they represent a meaningful objective that is difficult to achieve, but they are not the most representative indicator of the perceptual quality of a transcription system.

Figure 7 also shows that when the difference in $F_{\text{n,On}}$ is smaller than 10%, raters confidently disagree with $F_{\text{n,On}}$ as to which transcription is best nearly 40% of the time. This means that in these cases, $F_{\text{n,On}}$ should not be considered as a good descriptor of the quality of a transcription, at least from a perceptual point of view. This is particularly worrying, as very often, differences between systems are of the order of a few percentage points. On the other hand, we compare short segments, which means that a few errors could influence greatly $F_{\text{n,On}}$, while AMT systems are often compared over hours-long datasets. Also, in these difficult cases, raters tend to disagree more with each other, so personal judgement also comes into play. In summary, however, the majority of the previous analysis seems to indicate that $F_{\text{n,On}}$ is a good enough metric in clear-cut cases where the differences in performance are large, but should probably be treated with caution for small differences between AMT systems.

5. Defining a new metric

Given the relatively low agreement between ratings and current evaluation metrics, in particular in borderline cases, we propose to define a new evaluation metric, based on the ratings. The general idea is to compute a set of musical features on pairs (AMT output, target), and then train a classifier to output a value between 0 and 1 for each pair based on these features, using the ratings as training data.

5.1 Comments from participants

We first consider feedback from participants. Out of all participants, twelve left comments related to their decision-making strategies. The melody was mentioned as important in nine comments, making it the most important aspect according to comments, followed by rhythmic aspects (beat/meter/tempo, eight mentions) and harmony (four mentions). Some comments also mentioned higher level, less clearly defined aspects of music: three comments mentioned the "overall impression" was most important, two comments mentioned the presence of major artefacts or out-of-key notes. Overall, three comments mentioned explicitly that the presence of errors was not important as long as other aspects of the music were preserved, and most comments mentioned combinations of the above factors.

5.2 Feature description

From the previous comments, we define several features to capture various aspects of music, as well as typical AMT mistakes. In the following, we provide high-level definitions for each of these features. Full definitions can be found in the technical report accompanying this paper (Ycart et al., 2020).

5.2.1 Mistakes in highest and lowest voice

We use the highest and lowest voice of a piece as a proxy for the melody and the bassline, respectively. We define these metrics both framewise and notewise. For highest voice metrics, we define true positives and false negatives as notes in the highest voice of the target that have been correctly detected or missed (respectively). We count as a false positive any extra note that is above the highest voice in the target. From these values, we compute P, R, and F as described in Section 2.1. The lowest voice metrics are defined similarly. To better capture the score rather than the audio signal, we define the highest and lowest voices on targets without

taking the pedal into account, while the pedal is used in the computation of F_{f}, $F_{\mathrm{n,On}}$ and $F_{\mathrm{n,OnOff}}$.

5.2.2 Loudness of false negatives

We assume that missing a note that was loud in the original piece is more salient than missing a quiet one. We define two corresponding metrics:

- Average false negative loudness: the average MIDI velocity of false negatives. Each MIDI velocity is normalised by the average velocity in the ground truth in a two-second window centred on the false negative onset.
- False negative loudness ratio: the average ratio between the loudness of false negatives and the maximum loudness of active notes at the time of the false negative onset. We take into account the decay of long notes when computing the maximum loudness at the time of the onset.

5.2.3 Out-of-key false positives

We assume that out-of-key extra notes are much more noticeable than in-key ones. Instead of relying on key annotations, we define the key of a piece as the set of pitch classes that are active more than 10% of the time. The threshold of 10% is defined heuristically. This definition shows its limits when there are key modulations. We also define a non-binary key-disagreement as the proportion of the time that a pitch class is inactive. We then define two sets of metrics:

- Binary out-of-key: We count the number of false positives whose pitch is out-of-key. We then compute the proportion of out-of-key false positives among false positives, and among all notes in the output.
- Non-binary out-of-key: we compute the average key-disagreement of false positives, and the ratio between the sum of key-disagreements of false positives and the sum of key-disagreements of all detected notes.

5.2.4 Repeated and merged notes

A common type of mistake in AMT is to have repeated (i.e. fragmented) notes, or incorrectly merged notes. We count as a repeated note any false positive that overlaps with a ground-truth note of same pitch for at least 80% of its duration, and is preceded by at least one note of the same pitch that overlaps with the same ground-truth note. Conversely, we count as a merged note any false negative that overlaps for at least 80% of its duration with a detected note of same pitch and is preceded by at least one note of same pitch that overlaps with the same detected note. In both cases, we compute the proportion of mistakes among all false positives, and among all detected notes.

5.2.5 Specific pitch mistakes

It is also fairly common to have false positives in specific pitch intervals compared to ground-truth notes: semitone errors (neighbouring notes), octave errors (first partial), and 19 semitone errors (second partial). For these types of mistakes, we define both framewise and notewise metrics, for a given number of semitones n_s (here $n_s \in \{1, 12, 19\}$).

For framewise metrics, we count a specific pitch false positive for any false positive such that there is a ground truth note n_s semitones above or below. For notewise metrics, we count a specific pitch false positive for any false positive that overlaps for at least 80% of its duration with a ground truth note n_s semitones above or below. For $n_s = 19$, we only consider ground truth notes 19 semitones below, as second partial mistakes usually only happen 19 semitones above the ground truth. In both cases, we compute the proportion of mistakes among all false positives, and among all detected notes.

5.2.6 Polyphony level difference

We assume that a mistake is more salient when it is the only note being played and that it will also be noticeable if only a few notes of a big chord are transcribed. To account for this, we compute the absolute difference in polyphony level between the target and the output, at each timestep. We then use the mean, standard deviation, minimum and maximum values of this time series as features.

5.2.7 Rhythm histogram spectral flatness

Rhythm is another important aspect of music according to raters. We thus define a metric to account for rhythmic imprecision as follows. We first compute the inter-onset interval sequence of the output and the target. We keep simultaneous onsets, resulting in an IOI of 0. We then compute a histogram of the IOI values, with bin size of 10ms for IOIs below 100ms, and 100ms from 100ms to 2s (we drop IOIs above that value). This histogram should be more peaky for quantised MIDI files than outputs with rhythm imprecision. To describe this quantitatively, we compute the log-spectral flatness (Johnston, 1988) of both histograms (output and target). We use as a feature the spectral flatness of the output histogram, and the difference in spectral flatness between the output and target histograms.

5.2.8 Rhythm dispersion

We also propose another approach to characterising rhythm quality, based on K-means clustering (Murphy, 2012) of the IOI set. The general idea is to first run K-means clustering on the target IOIs, and then run K-means clustering on the output IOIs using the cluster centres of the target as initial values. We then compute the distance between cluster centres for the target and the output, as well as the relative difference in standard deviation within each cluster. We use as features the mean, maximum and minimum values across clusters.

Choosing the number of clusters is necessarily

heuristic. We determine the number of clusters by computing an IOI histogram as described in 5.2.7, but with wider bins, and choosing the peaks of that histogram as initial values for target IOI clustering.

5.3 Model fitting

Eventually, we aim to obtain a model that, given a set of features for a pair (AMT output, target), will output a scalar between 0 and 1. The main difficulty is that in our dataset, we do not have such absolute ratings, we only have pairwise comparison ratings. To achieve our goal, we draw inspiration from the contrastive loss approach (Hadsell et al., 2006). The original contrastive loss is defined as follows: given two inputs x_1 and x_2, a model f and a variable y such that $y = 1$ if x_1 and x_2 are considered similar, $y = 0$ otherwise:

$$L = y * |f(x_1) - f(x_2)|^2 + (1 - y)\max(\alpha - |f(x_1) - f(x_2)|, 0)^2 \quad (2)$$

In other words, if x_1 and x_2 are similar, the loss tries to bring their outputs together, and if they are dissimilar, it tries to push them apart. The α parameter is called the margin: if the distance between $f(x_1)$ and $f(x_2)$ is already greater than α, they are not moved further.

Given a target T, and two transcriptions of that target O_1 and O_2, we have, in place of x_1 and x_2, $g(T, O_1)$ corresponding to the set of features computed on T and O_1, and $g(T, O_2)$, the set of features computed on T and O_2. In our ratings, all transcriptions are dissimilar, so y is always equal to 0. Also, we do not only want $f(g(T, O_1))$ and $f(g(T, O_2))$ to be different, we also care about their order. We thus introduce a new variable z such that $z = 0$ if O_1 was chosen by the rater, and $z = 1$ if O_2 was chosen. We want to have $f(g(T, O_1)) > f(g(T, O_2))$ if $z = 0$, and the other way around if $z = 1$. We thus define our loss function as:

$$L = \max(\alpha - z * (f(x_2) - f(x_1)) - (1 - z)(f(x_1) - f(x_2)), 0)^2 \quad (3)$$

We incorporate the difficulty ratings in the margin: when ratings are confident, we use a higher margin. In practice, we use $\alpha = 0.5$ when Difficulty = 1, and decrease it by 0.1 for each difficulty level, until $\alpha = 0.1$ when Difficulty = 5.

We choose to use a simple model, allowing for interpretability of its parameters. Indeed, we want our metric to fit perceptual ratings, but also to serve as a diagnosis tool, allowing to easily investigate the contribution of each feature in the end result. For that reason, we use logistic regression, using as input all the above-defined features, in addition to the benchmark metrics.

5.4 Experiments

5.4.1 Setup

We use as input data to the logistic regression model the above features, along with the benchmark metrics defined in Section 2.1. We split our dataset between training, validation and test sets using a 90%-5%-5% partition, and use 20-fold cross-validation. The splits are made so there is no overlap in targets between the three subsets. There can be some overlap in terms of raters, which means that there is a possibility that the model learns the preferences of some specific participants. Our main concern is that the model should generalise to unseen input, so we still keep these ratings. In each fold, the data is z-normalised (mean=0 and variance=1). The weights of the logistic regression are all initialised to 0. The model is then trained using the Adam optimiser (Kingma and Ba, 2015) with a learning rate of 0.01 for a total of 3000 batches with a batch size of 100, which in practice is enough to ensure convergence. The parameters that achieve the lowest loss on the validation set are then used for testing. In each fold, we train 100 versions of the model (training a model takes about 15s), to account for potential variation in performance due to the randomness of the training process. We test whether our model agrees with ratings significantly better than $F_{\text{n,On}}$ by running an independent-samples T-test on each fold, and then testing whether the resulting T-values are significantly different from 0. We use 20 folds to have more data points when running the second test, and thus better statistical power in our results.

We focus the evaluation of our models on confident ratings. We thus compute the proportion of agreement between the output of our model and the confident ratings only, i.e. with Difficulty=1 or 2 (notated A_{conf}).

5.4.2 Results and ablation study

All results averaged across folds are shown in Figure 8. The dotted line corresponds to A_{conf} for $F_{\text{n,On}}$.

First, we train our model using all metrics. We manage to improve slightly (1%) but significantly ($p < 10^{-6}$) the agreement with the ratings, which is encouraging. It has to be noted that the model we used is very simple, and that more sophisticated models should be able to improve even further, though it may not be easy to achieve this without deteriorating interpretability.

In what follows, we investigate feature importance. One approach would be to inspect the weights of the trained logistic regression. However, it might happen that one feature has a high weight in a given model, but when removing it, its absence can be compensated by combinations of other features without decreasing performance. We thus favour an ablation approach to study how essential features are to model ratings, removing groups of features from the feature set and retraining our model as in Section 5.4.1. Table 6 summarises the configurations we investigate.

Three configurations perform significantly worse than All: NoFeatures, NoFramewise, and NoRhythm. Besides, NoFeatures is the only configuration that does not perform significantly better than $F_{\text{n,On}}$ ($p = 0.33$), which shows the usefulness of the feature set we have

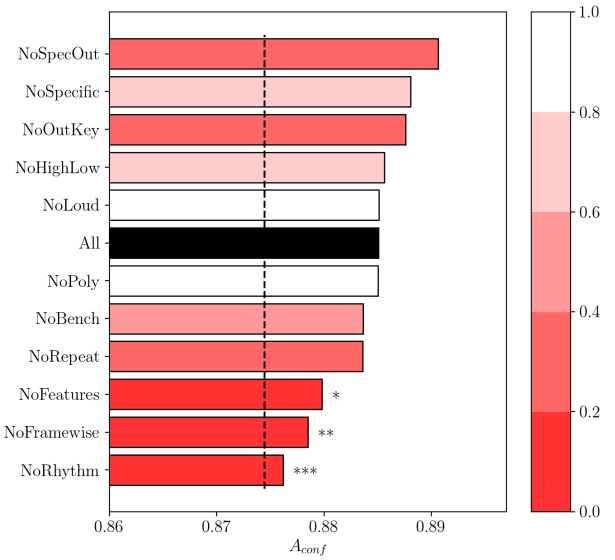

Figure 8: A_{conf} measure for each tested configuration, averaged across folds. The dotted line represents A_{conf} for $F_{\text{n,On}}$. Colors represent the p-value when testing whether each metric is different from the "All" configuration. Asterisks represent results significantly different from All (*: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$).

proposed. The low performance of NoRhythm compared to All shows the importance of the rhythm descriptors we used. This is somewhat contradictory with results from Section 4.3: we found that high tolerance thresholds for onsets and offsets gave better agreement, which seemed to indicate that temporal aspects are not important to raters. We suggest that our rhythm descriptor better captures higher-level aspects of rhythm reported as important to raters, such as the presence of a steady pulse and meter, rather than onset precision of individual notes. The fact that NoFramewise performs significantly worse than All shows that while F_{f} is indeed less correlated to ratings than $F_{\text{n,On}}$, some framewise metrics are useful and complementary to notewise metrics in modelling the ratings.

On the other hand, it appears that NoHighLow is not significantly worse than All. Yet, melody was the musical aspect that was most mentioned in user comments. We hypothesise that the reason this is not reflected in feature importance is that for the vast majority of examples in our dataset, the highest voice notewise F-measure, which best describes how well the melody was transcribed, is equal to 1. The model probably learns to give a low importance to that feature, as it is often constant. Another hypothesis is that our skyline approach to define the melody and the bassline might not correspond to perception. In the future, we might have to rely for instance on automatic melody estimation methods for symbolic music to better represent the melody.

Interestingly, it appears that some of the metrics we designed, in particular the out-of-key false positives and specific pitch errors, are actually counterproductive: removing them increases A_{conf}, although with relatively low significance ($p = 0.40$ and $p = 0.76$ respectively). We hypothesise that this is due to the definition of these metrics. For instance, if there are no specific pitch mistakes, this could either mean that there were no false positives (which is good), or there were a lot of false positives, none of which corresponded to a specific pitch (which is bad). This could lead to an interaction between specific pitch mistakes and benchmark precision metrics (e.g. penalise low specific pitch and low precision, but not low specific pitch and high precision). The same can be said of out-of-key false positives. However, such interactions cannot be represented by our model (simple logistic regression without interaction terms). As a result, out-of-key and specific pitch mistakes end up distracting the model more than they help. When removing both of these metrics (NoSpecOut configuration), our model reaches an A_{conf} of 89.1%. Removing other features that have either no impact or a negative impact on A_{conf} marginally decreases A_{conf} compared to NoSpecOut.

We make a pre-trained version of our metric available for future use (NoSpecOut configuration). We train it using all the data, without keeping out a validation or test set. Experiments show that in practice, the model does not overfit the training set: the training and validation losses are similar. We thus choose as final parameters those that minimise the loss over the whole training set. Given that we do not keep a held-out test set, we cannot report test performance of this specific released model.

6. Discussion

In this study, we presented a listening test to rate pairs of AMT systems. We compared perceptual ratings to results given by benchmark evaluation metrics. We have seen that most of the time, ratings agree with benchmark evaluation metrics, but in some cases (when both transcriptions have low $F_{\text{n,On}}$, and when the difference in $F_{\text{n,On}}$ between the two transcriptions is low), the agreement greatly decreases. We have proposed new quantitative measures describing musical features, and used them to define a new metric, that agrees with ratings significantly better than $F_{\text{n,On}}$. We also provide greater insight into which features were important to raters through an ablation study, illustrating in particular the importance of rhythm-related aspects.

Various aspects of this study could be improved. One of the most important would be to try more sophisticated models (e.g., artificial neural networks) to define a new metric. Indeed, the current approach only brings marginal improvement in A_{conf} compared to $F_{\text{n,On}}$, some more involved approaches could im-

Configuration	Removed features
All	None
NoBench	Benchmark metrics
NoFeatures	All features, except benchmark metrics
NoHighLow	Mistakes in highest and lowest voice
NoLoud	Loudness of false negatives
NoOutKey	Out-of-key false positives
NoRepeat	Repeated and merged notes
NoSpecific	Specific pitch mistakes
NoPoly	Polyphony level difference
NoRhythm	Rhythm histogram spectral flatness and rhythm dispersion
NoFramewise	Framewise benchmark metrics, framewise highest and lowest voice mistakes, framewise specific pitch errors, polyphony level difference, consonance measures
NoSpecout	Specific pitch mistakes and out-of-key false positives

Table 6: Description of each tested feature configuration.

prove further agreement with ratings. In particular, it would be theoretically possible to define a metric without using handcrafted features, directly by feeding the target and output into the system, but this approach would require more ratings to be trained robustly, and would lack interpretability. Still, some of the features might not have a linear influence on the quality of the transcription, and some may interact. Incorporating such factors into a model may improve performance. We chose a simple but interpretable logistic regression, which allowed us to verify easily the contribution of each metric to the final score.

Moreover, although we believe that absolute similarity rating between two excerpts is a difficult and ill-defined task (Allan et al., 2007; Flexer and Grill, 2016), it could be interesting to develop a listening test based on absolute similarity ratings between a reference and a single transcription. Provided inter-rater agreement is high enough, it would be interesting to train a regression model to approximate these ratings, and compare the results to those obtained with the current ranking paradigm.

Deeper investigation of the reasons for disagreement between ratings and $F_{\mathrm{n,On}}$ would also be useful to motivate the creation of new metrics. One way to investigate this would be to reproduce the above ablation study, but with a model trained and tested exclusively on ratings that disagree with $F_{\mathrm{n,On}}$, although the lack of data could make it difficult to achieve significant results, requiring collection of further ratings.

The generalisability of the metric we have designed should also be investigated. First, this metric was only designed for Western classical piano music. It would be interesting to investigate the extent to which it could be applied to other genres (e.g. jazz, non-Western music) and other instruments (e.g. guitar, multi-instrument ensembles). The protocol presented above could be applied with different stimuli to design metrics for other contexts, and potentially define a unified metric that works in every situation. But even in the context of Western classical piano music, some further experiments would have to be ran to test the generalisability of our metric. In particular, this metric was trained only on short segments; it remains to be seen whether it scales properly to longer pieces. One way to test our metric would be to run another similar listening test, once again using pairwise comparisons, but choosing specific, potentially artificial stimuli, to investigate specific points of disagreement: for instance, pairs of examples where our metric and $F_{\mathrm{n,On}}$ disagree as to which is best. By choosing representative examples with the specific aim of comparing these two metrics, much less data would be needed to validate which metric is most closely correlated to human perception.

Finally, this metric was designed to reflect perceptual similarity between the AMT output and the target. Such an evaluation criterion might not be relevant for every application. It is important when the overall musical quality of the transcription matters more than precise transcription of every note, for instance in the context of music creation and production (e.g. quick dictation of musical ideas) or tasks such as automatic accompaniment or cover detection. However, it might not be relevant in cases such as music education, where exact transcription of every note is paramount to properly assess the mistakes made by a student. In this case, reaching an $F_{\mathrm{n,OnOff}}$ of 1 should be the main objective, regardless of how the transcription sounds. In that regard, our metric complements the usual benchmark metrics to reflect perceptual quality of AMT outputs, but do not replace them.

7. Reproducibility

To allow further study of the data collected, we make it fully available, along with the stimuli, and the locations in seconds of the manually-selected cut points: `https://zenodo.org/record/3746863`

We also provide the code of the website: `https://github.com/adrienycart/AMT_perception_website`

A Python implementation of the used features and the pre-trained metric can be found here: `https://github.com/adrienycart/PEAMT`

Notes

[1] Soundfont download link: `http://freepats.zenvoid.org/Piano/acoustic-grand-piano.html`

Acknowledgements

The authors would like to thank Li Su and Tian Cheng for sharing their system implementations, and Rémi de Fleurian, Peter Harrison, Daniel Müllensiefen, Patrick E. Savage, Tillman Weyde, and Daniel Wolff for their useful suggestions on the design of this study.

AY is supported by a QMUL EECS Research Studentship. LL is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music and is supported by a China Scholarship Council and Queen Mary University of London joint PhD scholarship. EB is supported by UK RAEng Research Fellowship RF/128.

References

Allali, J., Ferraro, P., Hanna, P., and Robine, M. (2009). Polyphonic alignment algorithms for symbolic music retrieval. In *Auditory Display, 6th International Symposium, CMMR/ICAD*, pages 466–482.

Allan, H., Müllensiefen, D., and Wiggins, G. A. (2007). Methodological considerations in studies of musical similarity. In *Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR*, pages 473–478.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412.

Bay, M., Ehmann, A. F., and Downie, J. S. (2009). Evaluation of multiple-f0 estimation and tracking systems. In *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR*, pages 315–320.

Benetos, E., Dixon, S., Duan, Z., and Ewert, S. (2019). Automatic music transcription: An overview. *IEEE Signal Processing Magazine*, 36(1):20–30.

Bittner, R. M. and Bosch, J. J. (2019). Generalized metrics for single-f0 estimation evaluation. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR*, pages 738–745.

Cheng, T., Mauch, M., Benetos, E., and Dixon, S. (2016). An attack/decay model for piano transcription. In *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR*, pages 584–590.

Cogliati, A. and Duan, Z. (2017). A metric for music notation transcription accuracy. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR*, pages 407–413.

Daniel, A., Emiya, V., and David, B. (2008). Perceptually-based evaluation of the errors usually made when automatically transcribing music. In *Proceedings of the 9th International Conference on Music Information Retrieval, ISMIR*, pages 550–556.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer.

Emiya, V., Badeau, R., and David, B. (2010). Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech and Language Processing, TASLP*, 18(6):1643–1654.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Flexer, A. and Grill, T. (2016). The problem of limited inter-rater agreement in modelling music similarity. *Journal of new music research*, 45(3):239–251.

Hadsell, R., Chopra, S., and LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1735–1742.

Hawthorne, C., Elsen, E., Song, J., Roberts, A., Simon, I., Raffel, C., Engel, J. H., Oore, S., and Eck, D. (2018). Onsets and frames: Dual-objective piano transcription. In *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR*, pages 50–57.

Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C. A., Dieleman, S., Elsen, E., Engel, J. H., and Eck, D. (2019). Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *7th International Conference on Learning Representations, ICLR*.

Johnston, J. D. (1988). Transform coding of audio signals using perceptual noise criteria. *IEEE Journal on selected areas in communications*, 6(2):314–323.

Kelz, R., Dorfer, M., Korzeniowski, F., Böck, S., Arzt, A., and Widmer, G. (2016). On the potential of simple framewise approaches to piano transcription. In *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR*, pages 475–481.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR*.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

McLeod, A. and Steedman, M. (2018). Evaluating automatic polyphonic music transcription. In *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR*, pages 42–49.

Molina, E., Barbancho, A. M., Tardón, L. J., and Barbancho, I. (2014). Evaluation framework for automatic singing transcription. In *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*, pages 567–572.

Mongeau, M. and Sankoff, D. (1990). Comparison of

musical sequences. *Computers and the Humanities*, 24(3):161–175.

Müllensiefen, D., Gingras, B., Musil, J., and Stewart, L. (2014). The musicality of non-musicians: an index for assessing musical sophistication in the general population. *PloS one*, 9(2).

Müllensiefen, D., Gingras, B., Stewart, L., and Musil, J. (2011). The goldsmiths musical sophistication index (gold-msi): Technical report and documentation v1. 0. *London: Goldsmiths, University of London*.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., and Ellis, D. P. W. (2014). mir_eval: A transparent implementation of common MIR metrics. In *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*, pages 367–372.

Schramm, R., Nunes, H. D. S., and Jung, C. R. (2016). Audiovisual tool for solfège assessment. *ACM Transactions on Multimedia Computing, Communications, and Applications.*, 13(1).

Su, L. and Yang, Y.-H. (2015). Combining spectral and temporal representations for multipitch estimation of polyphonic music. *IEEE/ACM Transactions on Audio, Speech and Language Processing, TASLP*, 23(10):1600–1612.

Velardo, V., Vallati, M., and Jan, S. (2016). Symbolic melodic similarity: State of the art and future challenges. *Computer Music Journal*, 40(2):70–83.

Ycart, A. and Benetos, E. (2018). A-MAPS: Augmented MAPS dataset with rhythm and key annotations. In *ISMIR Late Breaking and Demos Papers*.

Ycart, A., Liu, L., Benetos, E., and Pearce, M. T. (2020). Musical features for automatic music transcription evaluation. Technical report, Queen Mary University of London, UK.